(2)

[illegible]MENTATION PAGE

Form Approved
OMB No. 0704-0188

[illegible]

1. AGENCY USE ONLY

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
FINAL/15 JAN 89 TO 14 OCT 92

4. TITLE AND SUBTITLE
FILTERING, STATISTICAL SIGNAL PROCESSING
& VARIATIONAL PROBLEMS (U)

5. FUNDING NUMBERS
2304/A6
61102F

6. AUTHOR(S)
Professor Sanjoy Mitter

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Institute of Technology
Informatioon & Decision Systems
Cambridge, MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER
AEOSR-TR- 93 0186

9. SPONSORING MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
110 DUNCAN AVE, SUITE B115
BOLLING AFB DC 20332-0001

10. SPONSORING MONITORING AGENCY REPORT NUMBER
AFOSR-89-0276

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

12b. DISTRIBUTION CODE
UL

13. ABSTRACT *(Maximum 200 words)*

During the grant period the PI made major contributions in three principal areas: (1) Robust Kalman filtering: (2) Structure determination for X-ray crystallography and (3) Stochastic recursive algorithms for global optimization. These theoretical advances have wide applications in diverse problems, such as identification of systems using maximum likelihood techniques, filtering in the presence of non-Gaussian observation noise, outlier detection, image analysis, and phase estimation problems.

14. SUBJECT TERMS

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR(SAME AS REPORT)

A Final Scientific Report of Research on Filtering, Statistical Signal Processing, and Variational Problems

under Grant No. 89-0276

for the period
15 January 1989 to 14 October 1992

submitted to

United States Air Force
Office of Scientific Research (AFOSR)
Bolling Air Force Base
Washington, D.C. 20032

(Attention: Dr. Jon A. Sjogren, Program Manager,
Probability, Statistics, and Signal Processing)

by Sanjoy K. Mitter
Principal Investigator

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

February 1993

1 Introduction

During the grant period 15 January 1989 to 14 October 1992, we have made major contributions in three principal areas:

- Robust Kalman filtering;
- Structure determination for X-ray crystallography; and
- Stochastic recursive algorithms for global optimization.

These theoretical advances have wide applications in diverse problems such as identification of systems using maximum likelihood techniques, filtering in the presence of non-Gaussian observation noise, outlier detection, image analysis, and phase estimation problems.

A technical overview of our research is presented in Section 2. This is followed by lists of the students, post-doctoral fellows, and faculty that have been supported by the grant, in Section 3; of invited presentations, in Section 4; and of publications based on the work described herein, in Section 5.

2 Description of Research

2.1 Robust Kalman Filtering

2.1.1 Introduction

Time-dependent data are often modeled by linear dynamic systems. Such representations assume that the data contain a deterministic component which may be described by a difference or differential equation. Deviations from this component are assumed to be random, and to have certain known distributional properties. These models may be used to estimate the "true" values of the data uncorrupted by measurement error, and possibly also to draw inference on the source generating the data.

Kalman Filtering has found an exceptionally broad range of applications, not only for estimating the state of a linear dynamic system in the presence of process and observation noise, but also for simultaneously estimating model parameters, choosing among several competing models, and detecting abrupt changes in the states, the parameters, or the form of the model. It is a remarkably versatile estimator, originally derived via orthogonal projections as a generalization of the Wiener filter to non-stationary processes, then shown to be optimal in a variety of settings: as the weighted least-squares solution to a regression problem, without regard to distributional assumptions; as the Bayes estimator assuming Gaussian noise, without regard to the cost functional; and as the solution to various game theoretic problems.

Neverthless, the Kalman Filter breaks down catastrophically in the presence of heavy-tailed noise, i.e. outliers. Even rare occurrences of unusually large observations severely degrade its performance, resulting in poor state estimates, non-white residuals, and invalid inference.

Statisticians and engineers often confront the problem of dealing with outliers in the course of model building and validation. Routinely ignoring unusual observations is neither wise, nor statistically sound, since such observations may contain valuable information as to unmodeled system characteristics, model degradation or breakdown, measurement errors, etc. But detecting unusual observations is only possible by comparison with the underlying trends and behavior; yet, it is precisely these that non-robust methods fail to capture when outliers are present. The purpose of robust estimators is thus twofold: to be as nearly optimal as possible when there are no outliers, i.e. under "normal" conditions; and to be resistent to outliers when they do occur, i.e. to be able to extract the underlying system behavior without being unduly affected by spurious values.

Past efforts to mitigate the effects of outliers on the Kalman Filter range from *ad hoc* practices such as simply discarding observations for which residuals are "too large," to more formal approaches based on non-parametric statistics, Bayesian methods, or minimax theory. An extensive survey of the literature is in [34, 35]. Many of these methods include heuristic approximations with ill-understood characteristics. While some have been empirically found to work well, their theoretical justifications have remained scanty at best. Their nonlinear forms, coupled with the difficulties inherent in dealing with non-normal distributions, have resulted in a strong preference in the literature for Monte Carlo simulations over analytical rigor.

In an effort to provide a more rigorous basis for sub-optimal filtering in the presence of non-Gaussian noise, a robust recursive estimator has been derived formally, combining Huber's theory of *minimax robust estimation* of a location parameter, recursive estimators based on the *stochastic approximation* theory of Robbins and Monro, and approximate *conditional mean estimation* based on *asymptotic expansion*. An overview of this approach appears in [32].

2.1.2 Preliminaries

Below, the notation $\mathcal{L}(\underline{x})$ denotes the probability law of the random vector $\underline{x}$ taking values in $\mathbf{R}^d$, $\mathcal{N}(\underline{\mu}, \Sigma)$ denotes a multivariate normal distribution with mean $\underline{\mu}$ and covariance Σ, and $\mathcal{N}(\underline{x}; \underline{\mu}, \Sigma)$ is the associated probability density function.

Consider first the model

$$\underline{z}_n = H_n \underline{\theta}_n + D_n \underline{v}_n, \tag{2.1}$$

where

$$\underline{\theta}_{n+1} = F_n \underline{\theta}_n + \underline{w}_n, \tag{2.2}$$

$n = 0, 1, \cdots$ denotes discrete time; $\underline{\theta}_n \in \mathbf{R}^q$ is the system state, with a random initial value distributed as $\mathcal{L}(\underline{\theta}_0) = \mathcal{N}(\bar{\underline{\theta}}_0, \Sigma_0)$; $\underline{z}_n \in \mathbf{R}^p$ is the observation (measurement); $\underline{w}_n \in \mathbf{R}^q$ is the process (plant) noise distributed as $\mathcal{L}(\underline{w}_n) = \mathcal{N}(\underline{0}, Q_n)$; $\underline{v}_n \in \mathbf{R}^p$ is the observation (measurement) noise distributed as $\mathcal{L}(\underline{v}_n) = \mathcal{F}$, a given distribution that admits a density and has mean and variance given by $\mathbf{E}[\underline{v}_n] = \underline{0}$ and $\mathbf{E}[\underline{v}_n \underline{v}_n^T] = R$; $\{F_n\}, \{H_n\}, \{D_n\}, \{Q_n\}$, Σ_0 and R are known matrices or sequences of matrices with appropriate dimensions; $\bar{\underline{\theta}}_0 \in \mathbf{R}^q$ is a known vector; and finally $\underline{\theta}_0$, $\underline{w}_n$, and $\underline{v}_n$ are mutually independent for all n.

The *Kalman Filter* is the estimator $\hat{\underline{\theta}}_n$ of the state $\underline{\theta}_n$ given the observations $\mathcal{Z}_n = \{\underline{z}_0, \cdots, \underline{z}_n\}$, and obeys the well-known recursion

$$\hat{\underline{\theta}}_n = \bar{\underline{\theta}}_n + K_n \underline{\gamma}_n, \tag{2.3}$$

where

$$\bar{\underline{\theta}}_n = F_{n-1} \hat{\underline{\theta}}_{n-1} \tag{2.4}$$

is the conditional *a priori* estimate of the state at time n (i.e., before updating by the observation $\underline{z}_n$) and

$$M_n = F_{n-1}P_{n-1}F_{n-1}^T + Q_{n-1} \tag{2.5}$$

is the conditional *a priori* estimation error covariance at time n.

$$\underline{\gamma}_n = \underline{z}_n - H_n\bar{\underline{\theta}}_n \tag{2.6}$$

is the innovation at time n and

$$\Gamma_n = H_nM_nH_n^T + D_nRD_n^T \tag{2.7}$$

is its covariance

$$K_n = M_nH_n^T\Gamma_n^{-1} \tag{2.8}$$

is the gain, and

$$P_n = M_n - K_n\Gamma_nK_n^T \tag{2.9}$$

is the *a posteriori* estimation error covariance at time n (i.e., after updating). The initial condition $\bar{\underline{\theta}}_0$ is given.

As is clear from Equations 2.3 and 2.6, the estimate is a linear function of the observation, a characteristic that is optimal only in the case of normally distributed noise or elliptical processes, which are sample-pathwise mixtures of normal processes. Similarly, Equations 2.5 and 2.8–2.9 show that the gain and covariance are independent of the data, a property related once again to the assumption of normality. Finally, the Gaussian case $\mathcal{F} = \mathcal{N}(\underline{0}, R)$, the residual (innovation) sequence $\{\underline{\gamma}_1, \cdots, \underline{\gamma}_n\}$ is white and is distributed as $\mathcal{L}(\underline{\gamma}_i) = \mathcal{N}(\underline{0}, \Gamma_i)$.

When $\mathcal{F}$ is a heavy-tailed distribution, on the other hand, the state estimation error can grow without bound (since the estimate is a linear function of the observation noise), the residual sequence becomes colored, and residuals become non-normal. Thus, not only is the estimate poor, but furthermore invalid inference would result from utilizing the residual sequence in the case of significant excursions from normality. A robust estimator should at the very least have the following characteristics: the state estimation error must remain bounded as a single observation outlier grows arbitrarily; the effect of a single observation outlier must not be spread out over time by the filter dynamics, i.e. a single outlier in the observation noise sequence must result in a single outlier in the residual sequence; and the residual sequence must remain nearly white when the observation noise is normally distributed except for an occasional outlier.

Such behavior could be obtained by replacing Equation 2.3 by, say,

$$\hat{\underline{\theta}}_n = \bar{\underline{\theta}}_n + K_n\underline{\psi}_n(\underline{\gamma}_n), \tag{2.10}$$

where $\underline{\psi}_n$ is an influence-bounding function that downweights "large" observations. In fact, a number of robust filters in the literature can be represented in the form 2.10. (See [35].) The significance of the functional ψ lies in the fact that it processes the innovation so as to mitigate the effects of observation outliers. "Overprocessing" the data results in loss of efficiency at the nominal model, while "underprocessing" makes the estimator excessively sensitive to outliers, i.e. non-robust. Some researchers have chosen ψ functions on the basis of engineering considerations, while others have derived them on probabilistic grounds, often using a Bayesian framework. The latter approach was taken here.

2.1.3 The Conditional Prior Distribution

Suppose the observation noise distribution $\mathcal{F}$ is a member of the ε-contaminated normal class of distributions

$$\mathcal{P}_{\varepsilon,R} = \{(1-\varepsilon)\mathcal{N}(0,R) + \varepsilon H : H \in \mathcal{S}\} \tag{2.11}$$

where $\mathcal{S}$ is the set of all suitably regular probability distributions, and $0 \leq \varepsilon \ll 1$ is the known fraction of "contamination." This form of the observation noise distribution can be used in an *asymptotic expansion*, in order to obtain a first-order approximation of the conditional prior distribution $\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1})$ of the state variable $\underline{\theta}_n$ given the observations $\mathcal{Z}_{n-1}$. A key property that ensures the finite dimensionality of this approximation is the *exponential stability* of the Kalman Filter, i.e. the fact that the effects of past observations decay fast enough. The resulting distribution is a perturbation from the normal, and all the pertinent parameters are given by various Kalman Filters and optimal smoothers that each make a specific assumption on the distribution of the noise at each point in time.

The first-order approximation of the conditional prior distribution $\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1})$ is next used to obtain a first-order approximation of the conditional mean of the state variable $\underline{\theta}_n$ given the observations $\mathcal{Z}_n$—i.e. to update the estimate by the current observation $\underline{z}_n$. This step uses a generalization of a proof due to [28, 29], made possible by a change in the order of integration. A similar derivation also yields the conditional covariance.

From 2.11, and assuming for now that $H = H^*$ is known, one can write

$$\underline{v}_n = (1-\eta_n)\underline{v}_n^{\mathcal{N}} + \eta_n \underline{v}_n^H \tag{2.12}$$

where η_n is a random variable independent of $\underline{\theta}_0$ and $\{\underline{w}_n\}$ obeying

$$\eta_n = \begin{cases} 0 & \text{w.p.} \quad (1-\varepsilon) \\ 1 & \text{w.p.} \quad \varepsilon \end{cases} \tag{2.13}$$

and $\{\underline{v}_n^{\mathcal{N}}\}$ and $\{\underline{v}_n^H\}$ are random variables independent of $\{\eta_n\}$, $\underline{\theta}_0$, and $\{\underline{w}_n\}$ with $\mathcal{L}(\underline{v}_n^{\mathcal{N}}) = \mathcal{N}(\underline{0}, R)$ (for some $R > 0$) and $\mathcal{L}(\underline{v}_n^H) = H^*$. Finally, loosely defining a random variable distributed as H^* as an "outlier," denote the event "there has been no outlier among the first n observations" by $\mathcal{H}_n = \{\eta_0 = 0, \cdots, \eta_n = 0\}$, and the event "there has been exactly one outlier among the first n observations, at time $i-1$" by $\mathcal{H}_n^i = \{\eta_0 = 0, \cdots, \eta_{i-2} = 0,\ \eta_{i-1} = 1,\ \eta_i = 0, \cdots, \eta_n = 0\}$. Then, it is easy to verify that

$$\begin{aligned} &\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1})\mathbf{p}(\mathcal{Z}_{n-1}) \\ &\quad = \mathbf{p}(\mathcal{H}_{n-1})\mathbf{p}(\mathcal{Z}_{n-1}|\mathcal{H}_{n-1})\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1},\mathcal{H}_{n-1}) \\ &\quad + \sum_{i=1}^{n} \mathbf{p}(\mathcal{H}_{n-1}^i)\mathbf{p}(\mathcal{Z}_{n-1}|\mathcal{H}_{n-1}^i)\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1},\mathcal{H}_{n-1}^i) \\ &\quad + \text{higher-order terms.} \end{aligned} \tag{2.14}$$

Clearly, the first term on the right-hand side of 2.14 is the distribution conditioned on the event that there were no outliers, each term in the summation to the event that there was exactly one outlier, and the higher-order terms to the occurrence of two or more outliers. Moreover, defining $\mathcal{Z}_n^i = \{\underline{z}_0, \cdots, \underline{z}_{i-2}, \underline{z}_i, \cdots, \underline{z}_{n-1}\}$, it follows that

$$\begin{aligned} &\mathbf{p}(\mathcal{Z}_{n-1}|\mathcal{H}_{n-1}^i)\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1}, \mathcal{H}_{n-1}^i) \\ &\quad = \mathbf{p}(\mathcal{Z}_{n-1}^i|\mathcal{H}_{n-1}^i)\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1}^i, \mathcal{H}_{n-1}^i)\mathbf{p}(\underline{z}_{i-1}|\underline{\theta}_n, \mathcal{Z}_{n-1}^i, \mathcal{H}_{n-1}^i). \end{aligned} \tag{2.15}$$

Note that the *only* non-normal term on the right-hand side of 2.15 is the last one. All other terms in 2.15, as well as in 2.14, are normal. These remarks are formalized in the following theorem. Note first that if the system is completely observable and completely controllable, then given any $\tilde{\underline{\theta}}_0 < \infty$, and defining the closed-loop recursion

$$\tilde{\underline{\theta}}_{n+1} = (I - K_{n+1}H_{n+1})F_n\tilde{\underline{\theta}}_n, \tag{2.16}$$

there exist $\lambda > 0$ and $0 < \delta < 1$ such that

$$||\tilde{\underline{\theta}}_n|| < \lambda\delta^n \tag{2.17}$$

i.e. the filter is exponentially asymptotically stable.

Theorem 2.1.1 *Let the system given by Equations 2.1-2.2 be stable, and let δ be a real number for which 2.17 holds. Let ω be the smallest integer such that*

$$\delta^\omega \leq \varepsilon \tag{2.18}$$

If

$$\omega\varepsilon < 1 \tag{2.19}$$

and if the distribution H^* has bounded moments, then

$$\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1}) = (1-\varepsilon)^\omega \kappa_n \kappa_n^0 \mathcal{N}(\underline{\theta}_n; \bar{\underline{\theta}}_n^0, M_n^0) \tag{2.20}$$

$$+ \ \varepsilon(1-\varepsilon)^{\omega-1}\kappa_n \sum_{i=n-\omega+1}^{n} \kappa_n^i \mathcal{N}(\underline{\theta}_n; \bar{\underline{\theta}}_n^i, M_n^i) \tag{2.21}$$

$$\int \mathcal{N}(\underline{z}_{i-1} - \underline{\xi}; H_{i-1}\underline{\nu}_n^i + H_{i-1}V_n^i(\underline{\theta}_n - \underline{\theta}_n^i), \tag{2.22}$$

$$H_{i-1}W_n^iH_{i-1}^T - H_{i-1}V_n^iM_n^iV_n^{iT}H_{i-1}^T)dH^*(\underline{\xi}) \tag{2.23}$$

$$+ \ O_p(\omega^2\varepsilon^2) \tag{2.24}$$

for all $n \geq \omega$, where, for $i = 0, 1, \cdots$ and $n > i$,

$$\bar{\underline{\theta}}_n^i = F_{n-1}\underline{\theta}_{n-1}^i \tag{2.25}$$

$$\underline{\theta}_n^i = \bar{\underline{\theta}}_n^i + K_n^i\underline{\gamma}_n^i \tag{2.26}$$

$$M_n^i = F_{n-1}P_{n-1}^iF_{n-1}^T + Q_{n-1} \tag{2.27}$$

$$\underline{\gamma}_n^i = \underline{z}_n - H_n\bar{\underline{\theta}}_n^i \tag{2.28}$$

$$\Gamma_n^i = H_nM_n^iH_n^T + D_nRD_n^T \tag{2.29}$$

$$K_n^i = M_{n-1}^i H_{n-1}^T {\Gamma_{n-1}^i}^{-1} \tag{2.30}$$

$$P_n^i = M_n^i - K_n^i \Gamma_n^i {K_n^i}^T \tag{2.31}$$

and

$$\kappa_n^i = \kappa_{n-1}^i \mathcal{N}(\underline{\gamma}_{n-1}^i; \underline{0}, \Gamma_{n-1}^i), \tag{2.32}$$

for $1 = 1, 2, \cdots$ *and* $n > i$.

$$V_n^i = V_{n-1}^i P_{n-1}^i F_{n-1}^T {M_n^i}^{-1} \tag{2.33}$$

$$\underline{\nu}_n^i = \underline{\nu}_{n-1}^i + V_{n-1}^i K_{n-1}^i \underline{\gamma}_{n-1}^i \tag{2.34}$$

$$W_n^i = W_{n-1}^i - V_{n-1}^i K_n^i \Gamma_n^i {K_n^i}^T {V_{n-1}^i}^T, \tag{2.35}$$

subject to the initial conditions

$$\bar{\underline{\theta}}_i^i = F_{i-1} \bar{\underline{\theta}}_{i-1}^0 \tag{2.36}$$

$$M_i^i = F_{i-1} M_{i-1}^0 F_{i-1}^T + Q_{i-1} \tag{2.37}$$

$$V_i^i = M_{i-1}^0 F_{i-1}^T {M_i^i}^{-1} \tag{2.38}$$

$$\underline{\nu}_i^i = \bar{\underline{\theta}}_{i-1}^0 \tag{2.39}$$

$$W_i^i = M_{i-1}^0 \tag{2.40}$$

$$\kappa_i^i = \kappa_{i-1}^0 \tag{2.41}$$

for $i > 0$, *and*

$$\underline{\theta}_0^0 = \bar{\underline{\theta}}_0 \tag{2.42}$$

$$M_0^0 = M_0 \tag{2.43}$$

$$\kappa_0^0 = 1. \tag{2.44}$$

The normalization constant satisfies

$$\kappa_n^{-1} = (1-\varepsilon)^{\omega} \kappa_n^0 \tag{2.45}$$

$$+\varepsilon(1-\varepsilon)^{\omega-1} \sum_{i=n-\omega+1}^{n} \kappa_n^i \int \mathcal{N}(\underline{z}_{i-1} - \underline{\xi}; H_{i-1}\underline{\nu}_n^i, \tag{2.46}$$

$$H_{i-1} W_n^i H_{i-1}^T) dH^*(\underline{\xi}) \tag{2.47}$$

(The case $n < \omega$ *is very similar.)*

Proof. See [34, 35]. ■

Note that Equations 2.25–2.31 are a bank of Kalman Filters, each starting at a different point in time $i = 0, 1, 2, \cdots$. Because of the way in which they are initialized, the cases $i > 0$ correspond to Kalman Filters skipping the $i-1$st observation. The case $i = 0$ is based on all observations. Similarly, Equations 2.33–2.35 are a bank of optimal fixed-point smoothers, each estimating the state at a different point in time $i = 0, 1, 2, \cdots$, based on all preceeding and subsequent observations. Thus, each term in the summation on the right-hand side of 2.24 is a Kalman Filter that skips one observation, coupled with an optimal smoother that estimates the state at the time the observation is skipped.

Evidently, as $n \to \infty$, the probability of the event that only a finite number of outliers occur vanishes for any $\varepsilon > 0$. That the density can nevertheless be approximated by the first-order expression in 2.24 is due to the exponential asymptotic stability of the Kalman Filter: ω represents a "window size" beyond which the effects of older observations have sufficiently attenuated.

2.1.4 The Conditional Mean Estimator

The approximate conditional prior distribution $\mathbf{p}(\underline{\theta}_n|\mathcal{Z}_{n-1})$ of Theorem 2.1.1 is now used to derive the conditional mean and variance, respectively denoted by

$$\underline{T}_n = \mathbf{E}[\underline{\theta}_n|\mathcal{Z}_n] \tag{2.48}$$

and

$$\Sigma_n = \mathbf{E}[(\underline{\theta}_n - \underline{T}_n)(\underline{\theta}_n - \underline{T}_n)^T|\mathcal{Z}_n]. \tag{2.49}$$

Let h^* denote the density associated with H^*, provided that it exists.

Theorem 2.1.2 *Let the conditions of Theorem 2.1.1 be satisfied for the system given by Equations 2.1–2.2. If h^* exists and is bounded and differentiable a.e., then*

$$\underline{T}_n = (1-\varepsilon)^\omega \kappa_{n+1} \pi_n^0 \underline{T}_n^0 + \varepsilon(1-\varepsilon)^{\omega-1}\kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \underline{T}_n^i + O_p(\omega^2\varepsilon^2) \tag{2.50}$$

for all $n \geq \omega$, where

$$\underline{T}_n^0 = \bar{\underline{\theta}}_n^0 + M_n^0 H_n^T \underline{\psi}_n^0(\underline{z}_n - H_n \bar{\underline{\theta}}_n^0) \tag{2.51}$$

$$\underline{T}_n^i = \underline{\theta}_n^i + P_n^i V_n^{iT} H_{i-1}^T \underline{\psi}_n^i(\underline{z}_{i-1} - H_{i-1}\underline{\nu}_{n+1}^i) \tag{2.52}$$

$$\pi_n^0 = (1-\varepsilon)\kappa_{n+1}^0 + \varepsilon\kappa_n^0 \int \mathcal{N}(\underline{z}_n - \underline{\xi}; H_n\underline{\theta}_n^0, H_n M_n^0 H_n^T) h^*(\underline{\xi}) d\underline{\xi} \tag{2.53}$$

$$\pi_n^i = (1-\varepsilon)\kappa_{n+1}^i \int \mathcal{N}(\underline{z}_{i-1} - \underline{\xi}; H_{i-1}\underline{\nu}_{n+1}^i, H_{i-1} W_{n+1}^i H_{i-1}^T) h^*(\underline{\xi}) d\underline{\xi} \tag{2.54}$$

and the influence-bounding functions are given by

$$\underline{\psi}_n^0(\underline{\zeta}) = -\frac{\underline{\nabla}_\zeta \left((1-\varepsilon)\mathcal{N}(\underline{\zeta}; \underline{0}, \Gamma_n^0) + \varepsilon \int \mathcal{N}(\underline{\zeta} - \underline{\xi}; \underline{0}, H_n M_n^0 H_n^T) h^*(\underline{\xi}) d\underline{\xi}\right)}{(1-\varepsilon)\mathcal{N}(\underline{\zeta}; \underline{0}, \Gamma_n^0) + \varepsilon \int \mathcal{N}(\underline{\zeta} - \underline{\xi}; \underline{0}, H_n M_n^0 H_n^T) h^*(\underline{\xi}) d\underline{\xi}} \tag{2.55}$$

$$\underline{\psi}_n^i(\underline{\zeta}) = -\frac{\underline{\nabla}_\zeta \left(\int \mathcal{N}(\underline{\zeta} - \underline{\xi}; \underline{0}, H_{i-1} W_{n+1}^i H_{i-1}) h^*(\underline{\xi}) d\underline{\xi}\right)}{\int \mathcal{N}(\underline{\zeta} - \underline{\xi}; \underline{0}, H_{i-1} W_{n+1}^i H_{i-1}) h^*(\underline{\xi}) d\underline{\xi}} \tag{2.56}$$

with $\bar{\underline{\theta}}_n^i, \underline{\gamma}_n^i, K_n^i, M_n^i, P_n^i, \Gamma_n^i, V_n^i, \underline{\nu}_n^i, W_n^i, \kappa_n^i$, and κ_n as defined in Theorem 2.1.1, subject to the same initial conditions. Furthermore,

$$\Sigma_n = (1-\varepsilon)^\omega \kappa_{n+1} \pi_n^0 \Sigma_n^0 + \varepsilon(1-\varepsilon)^{\omega-1}\kappa_{n+1} \sum_{i=n-\omega+1}^{n} \pi_n^i \Sigma_n^i + O_p(\omega^2\varepsilon^2) \tag{2.57}$$

for all $n \geq \omega$, where

$$\underline{\Sigma}_n^0 = M_n^0 - M_n^0 H_n^T \Psi_n^0(\underline{z}_n - H_n \underline{\bar{\theta}}_n^0) H_n M_n^0 + (\underline{T}_n - \underline{T}_n^0)(\underline{T}_n - \underline{T}_n^0)^T \tag{2.58}$$

$$\underline{\Sigma}_n^\iota = P_n^\iota - P_n^\iota {V_n^\iota}^T H_{\iota-1}^T \Psi_n^\iota(\underline{z}_{\iota-1} - H_{\iota-1} \underline{\nu}_{n+1}^\iota) H_{\iota-1} V_n^\iota P_n^\iota + (\underline{T}_n - \underline{T}_n^\iota)(\underline{T}_n - \underline{T}_n^\iota)^T. \tag{2.59}$$

and Ψ_n^ι is given by

$$\Psi_n^\iota(\underline{\zeta}) = \underline{\nabla}_{\underline{\zeta}} {\underline{\psi}_n^\iota}^T(\underline{\zeta}). \tag{2.60}$$

(The case $n < \omega$ is very similar.)

Proof. See [34, 35]. ∎

Both Theorem 2.1.1 and Theorem 2.1.2 are based on the assumption that outliers occur rarely relative to the dynamics of the filter. In the unlikely event that two outliers occur within less than ω time steps of each other, Equation 2.52—which shows that $\underline{T}_n^\iota$ is linear in $\underline{z}_n$ —suggests that the estimate would be strongly affected. This implies that the estimator developed here is robust in the presence of rare and isolated outliers, but not when outliers occur in batches.

The estimator is a weighted sum of *stochastic approximation*-like estimators, with weights equal to the posterior probabilities of each outlier configuration. These probabilities are conditioned on all the observations, including the current one. Since the banks of parallel filters and smoothers are entirely independent of each other, the estimate derived here is well suited to parallel computation. Furthermore, the covariance is a function of a set of matrices $\{M_n^\iota\}$, $\{P_n^\iota\}$, $\{\Gamma_n^\iota\}$, $\{V_n^\iota\}$, and $\{W_n^\iota\}$, which are themselves independent of the observations. Thus, they can be pre-computed and stored, as is sometimes done with the Kalman Filter. Although the covariance given by 2.57 is not independent of the data (a feature that would only be optimal in the normal case), this implies that a great deal of computation may nevertheless be performed off-line.

Finally, it is easy to verify that, for $\varepsilon = 0$,

$$\begin{aligned} \underline{\psi}_n^0(\underline{\gamma}_n^0) &= -\frac{\underline{\nabla}_\gamma \mathcal{N}(\underline{\gamma}; \underline{0}; \Gamma_n^0)|_{\gamma=\underline{\gamma}_n^0}}{\mathcal{N}(\underline{\gamma}_n^0; \underline{0}, \Gamma_n^0)} & (2.61) \\ &= {\Gamma_n^0}^{-1} \underline{\gamma}_n^0, & (2.62) \end{aligned}$$

so that $\underline{T}_n$ reduces to the Kalman Filter when the noise is normally distributed.

2.2 X-Ray Crystallography

2.2.1 Introduction

A new Markov random field-based algorithm has been proposed for signal reconstruction from Fourier transform magnitude motivated by the data reduction calculations of X-ray crystallography [12, 5, 11, 14, 8, 6, 7, 9, 13, 15, 10]. The purpose of an X-ray crystallography experiment is to determine the position in three dimensional space of each atom in a molecule. The measured data are the magnitudes squared of the Fourier transform of the electron density function of a crystal of the molecule of interest and possibly also of chemical derivatives. The data reduction calculations are a signal reconstruction problem for the three dimensional electron density. In the so-called "direct" methods of interest here, the reconstruction is based on a noisy measurement of the magnitude squared of the Fourier transform of the electron density of a single crystal, that is, no chemical derivatives of the molecule are studied.

These reconstruction problems are unusual [30, 24]. For instance, it is the periodicity of the crystal that samples the Fourier transform of the three dimensional repeat unit (called a "unit cell"), so that the sampling is beyond the control of the investigator, and the sampling rate is below the Nyquist rate for the autocorrelation function that can be computed from the available Fourier transform magnitudes. Furthermore, the electron density is invariant under a space group symmetry.

The most powerful direct methods are probabilistic in nature [20, 1, 2], are based on a model in which the atomic locations are independent random variables, and are successful on small molecules. The failure of these techniques to extend to larger molecules is attributed by Bricogne [1, 2] to inconsistent usage of probabilistic information and inaccurate computation of marginal probabilities. In addition, he notes the very idealized nature of the standard independent atomic location hypothesis.

There are three major themes in the work reported here: tractable incorporation of *a priori* information, consistent use of probabilistic information, and analytical (rather than numerical) approximations. The starting point is a Markov random field (MRF) *a priori* model for the electron density; a Bayesian statistical problem whose solution is the thresholded conditional mean of the MRF given the data is defined; and the conditional mean is approximately computed using symmetry breaking, the spherical model, and small noise asymptotics. Initial results from work at MIT are reported in [5, 8, 6, 7] and further results from work continued at Purdue University (School of Electrical Engineering) are described in [12, 11, 14, 9, 13, 15, 10].

2.2.2 The MRF *a priori* Model and the *a posteriori* Model

The MRF defines a probability distribution on a collection of binary random variables $\phi_{\vec{n}} \in \{0, 1\}$ which lie on a lattice. The connection between the MRF and the electron density is that the atoms are restricted to lie on the lattice and site $\vec{n}$ is occupied by an atom if and only if $\phi_{\vec{n}} = 1$. This construction assures a positive and atomic electron density. It remains to arrange the correct spacing between atoms, which is achieved by the Hamiltonian H^{apriori} of the MRF. H^{apriori} is the sum of energies $u_{\vec{n}}$ associated with each site in the lattice. The idea behind $u_{\vec{n}}$ is simple: If an atom is not present at site $\vec{n}$ then $u_{\vec{n}} = 0$. If an atom is present at site $\vec{n}$ then

1. if other atoms are located within a minimum bond radius of length r_1 then $u_{\vec{n}}$ is positive because the atoms are unphysically close while

2. if no other atoms are located within a maximum bond radius of length r_2 then $u_{\vec{n}}$ is positive because the atom at site $\vec{n}$ is floating free unbound in the molecule while

3. if one or more atoms are located between the minimum and maximum bond radii and none are located closer than the minimum bond radius then $u_{\vec{n}}$ is negative because the atom at site $\vec{n}$ can be correctly bound.

While a variety of complicated functional forms can be chosen for $u_{\vec{n}}$, good success has been achieved with quadratic forms which make possible a wide range of analytic calculations. Note that H^{apriori} is invariant under translations, rotations, and inversions of the field ϕ.

In light of the relationship between ϕ and the electron density, the exact observations are the magnitude squared of the Fourier coefficients of ϕ. The actual data $y_{\vec{k}}$ are additively corrupted by noise which is modeled as Gaussian with zero mean and known $\vec{k}$-dependent variance $\sigma^2_{\vec{k}}$.

The joint and *a posteriori* distributions on ϕ and y can be written as MRFs so the calculation of the conditional mean is simply the calculation of the spatially varying mean of this new MRF. The Hamiltonian for this MRF is $H^{\text{apriori}} + H^{\text{obs}}$ where H^{obs} comes from the Gaussian conditional observation distribution. There is, however, a problem. Specifically, the invariance of H^{apriori} under translations, rotations, and inversions of the field ϕ and the lack of phase measurements implies that the mean of the new MRF is a constant. In order to solve this problem the Hamiltonian is modified by introducing a symmetry breaking term $H^{\text{s.b.}}$ which is proportional to $\sum_{\vec{n}} \psi_{-\vec{n}}\phi_{\vec{n}}$. This is a good choice because for suitable ψ, called the "kernel", it breaks the symmetries and because it is linear and can thus be viewed as a small perturbation. The values of the variables $\{\psi_{\vec{n}}\}$ are set by a data adaptive optimization described below.

2.2.3 Bayesian Estimation and Computation of the Conditional Mean

The cost that is minimized in order to derive the Bayesian estimator is the mean squared error between the true and reconstructed fields. For these binary fields, the "segmentation" cost that applies an equal penalty to any reconstruction error leads to the same estimator. The result of the minimization problem is that the estimator has two steps: first compute the conditional mean of the electron density ϕ given the data y and then threshold the result at value 1/2 so that sites with conditional mean greater (less) than 1/2 take value 1 (0). As mentioned above, the conditional mean is computed by computing the spatially varying mean of the new MRF which has Hamiltonian $H^{\text{apriori}} + H^{\text{obs}} + H^{\text{s.b.}}$. This calculation is done through two approximations: First, the spherical model is introduced in order to relax the $\phi_{\vec{n}} \in \{0,1\}$ constraint. It transforms a sum over the corners of a hypercube into an integral over the surface of a hypersphere inscribed around the hypercube. Half of the integrations can be done analytically but the remaining half are intractable exponential-of-quartic integrations. Therefore the second approximation is made which is the evaluation of these integrals by small-noise asymptotic techniques where the "small noise" refers to small observation noise, i.e., small $\sigma^2_{\vec{k}}$. (This is the relevant limit in X-ray crystallography). The key step in the asymptotics is the calculation of the critical point (i.e., the global minimum of the exponent), which can be done exactly with computation linear in the size of the MRF lattice. The results of these two approximations are analytic formulas for the conditional mean of the Fourier coefficients of the field given as functions of the critical point and the kernel ψ of the symmetry breaking.

2.2.4 Data Adaptation

The kernel ψ is chosen to minimize a cost function of the conditional mean of the field ϕ given the data y. This optimization makes the estimator adapt to the data. The primary purpose of the adaptation is to ameliorate the errors introduced by the spherical model. The cost penalizes excursions of the mean outside of the interval [0.1] (which are exclusively due to the approximations since $\varphi_{\vec{n}} \in \{0, 1\}$), penalizes excursions from the two endpoints 0 and 1 (since one desires a ψ that results in a confident estimator), and penalizes deviations of the energy in ψ from a target (since one does not want ψ to vanish and hence fail to break the symmetries or to grow too large so that $H^{\mathrm{s.b.}}$ dominates the total Hamiltonian).

Once ψ is chosen the conditional mean of the Fourier coefficients of the field ϕ can be calculated. Then the Fourier series is inverted to compute the conditional mean of the field ϕ. Finally, the conditional mean is thresholded at value 1/2, that is, an atom is placed at each lattice site where the conditional mean exceeds 1/2. One and two dimensional numerical examples are given in [6].

2.2.5 Incorporation of Space Group Symmetries

The unit cell is the periodic repeat unit of the crystal. The presence of a nontrivial space group symmetry means that there is additional structure within the unit cell. For example, the unit cell might be divided in half with the electron density in one half the mirror image of the electron density in the other half. The space group is known before the reconstruction is done. In one dimension there are only two space groups: the trivial group $P1$ where there is no structure within the unit cell (i.e., a periodic function) and the group $P\bar{1}$ for which there is a mirror point of symmetry in the middle of the unit cell (i.e., periodic and even). In three dimensions the situation is much more complicated and there are a total of 230 space groups [21].

Three approaches to solving signal reconstruction problems in the presence of nontrivial space groups are described [12, 11, 14, 15, 10]. In Approach 1, the actual space group $\mathcal{G}$ is replaced by the subgroup $P1$, the signal reconstruction results of [8, 6] are applied, and then the invariance under $\mathcal{G}$ information is added in two ways. First, reconstructions that are invariant under $P1$ but not $\mathcal{G}$ are transformed into reconstructions invariant under $\mathcal{G}$ by averaging. Second, the invariance of the signal under $\mathcal{G}$ is applied as a soft constraint by adding a term to the C cost function for ψ optimization. The advantage of Approach 1 is simplicity since Ref. [8, 6] is applied with little alteration to any space group $\mathcal{G}$. The disadvantage is the suboptimal use of space group information. Furthermore, the data adaptation–minimization of C with respect to ψ–occurs in a higher dimensional space than is necessary. Symmetry breaking is retained.

The second and third approaches both integrate the presence of the space group $\mathcal{G}$ as a hard constraint into the signal reconstruction process. The two approaches differ by the order in which noncommuting nonlinear operations are performed: in Approach 2 the spherical model is applied before the space group symmetry is enforced (so that the spherical model i^ applied to the entire unit cell) while in Approach 3 the order is reversed (so that the spherical model is applied only to the asymmetric unit). (The asymmetric unit is a minimum subset of the unit cell that is sufficient to determine the electron density in the entire unit cell). The advantage of Approach 2 is that the calculation of the critical point in the small observation noise asymptotics is only slightly changed from Refs. [8][6, Appendix A]. Therefore it can be done analytically. The disadvantage is that the spherical model approximation is applied over a larger number of sites (the entire unit cell) and so it is less accurate. Symmetry breaking is required. The advantage of Approach 3 is that the spherical model is applied over a smaller number of sites (only the asymmetric unit) and so it is more accurate. The disadvantage is that the calculation of the critical point in the small observation noise asymptotics is substantially more difficult than :n Refs. [8][6, Appendix A] and to date an analytical solution is available only for a special case. Symmetry breaking is not required, mirroring the fact that symmetry breaking is not required in an exact solution. In fact, if used, symmetry breaking only influences the value of second and higher order terms in the asymptotic expansion. In both Approaches 2 and 3 the data adaptation occurs in the smallest possible dimensional space.

The methods are compared [12, 11, 14, 15, 10] in 1D for space group $P\bar{1}$. Figure 1 shows performance, measured as $E\sqrt{\sum_n(\phi_n - \hat{\phi}_n)^2}$, for six different estimators as a function of the observation noise standard deviation σ for a $L = 17$ sites lattice. All results are Monte Carlo computations using 1000 realizations. The dashed lines E and A are estimators from Refs. [13, 8, 6] which are unaware of the presence of $P\bar{1}$ symmetry. E is the exact estimator computed by explicitly summing over all ϕ configurations. Symmetry breaking is present. This estimator is totally impractical for any reasonable sized lattice and is the reason for the choice of $L = 17$ for these simulations. However, it is the optimal Bayesian estimator in the absence of space group information. A is the approximate estimator from Refs. [13, 8, 6]. The solid lines are estimators that are aware of the presence of $P\bar{1}$ symmetry. E^* is the exact estimator computed by explicitly summing over the $P\bar{1}$ symmetric subset of ϕ configurations (but with symmetry breaking turned off). $A1$, $A2$, and $A3$ are the Approach 1, Approach 2, and Approach 3 estimators. The critical point for the small observation noise asymptotics for Approach 3 was determined numerically by N2ONG (Ref. [22, Section 8.4, pp. 903-908]).

Note several aspects of these numerical results: Knowledge that the signal is $P\bar{1}$ symmetric is very valuable–compare E with E^*; with knowledge that the signal is $P\bar{1}$ symmetric, symmetry breaking is not required–see E^*; Approaches 1 and 2 provide roughly equivalent performance, performance that sometimes exceeds that of the optimal estimator E that is unaware of the $P\bar{1}$ symmetry (and is very expensive to compute); and Approach 3 provides poor performance which is attributed to the lack of data adaptation.

2.2.6 Analytical Gradients for Data Adaptation Optimization

In the cited work, the optimization of ψ was done using a multidimensional downhill simplex method [33, Section 10.4 pp. 305-309]. Evaluating the cost function requires two FFTs. A natural improvement is to use a conjugate gradient algorithm with analytical gradients. The fact that the gradient can be computed analytically is not surprising though the calculation requires care because, for example, ψ is real so that Ψ is conjugate symmetric. What is surprising is that the cost function and its complete gradient can be computed at a cost of four FFTs–only twice as much computation as was required for the function value alone.

The algorithm for efficient gradient calculation has been worked out in 3D for an Approach 2 estimator for the monoclinic $C2$ space group. (The equations are not included here). Results using a Fletcher-Reeves-Polak-Ribiere conjugate gradient algorithm [33, Section 10.6 pp. 318-322] on a $4 \times 4 \times 4$ problem with a 2% observation noise standard deviation (realistic for small molecule X-ray crystallography) are shown in Figure 2. The z axis is $E\sqrt{\sum_n(\phi_n - \hat{\phi}_n)^2}$ and the x and y coordinates are two parameters in the cost function C. Note both the excellent performance achieved and the relative insensitivity of the performance to the values of the two parameters. Extension of these results to experimental data is currently underway.

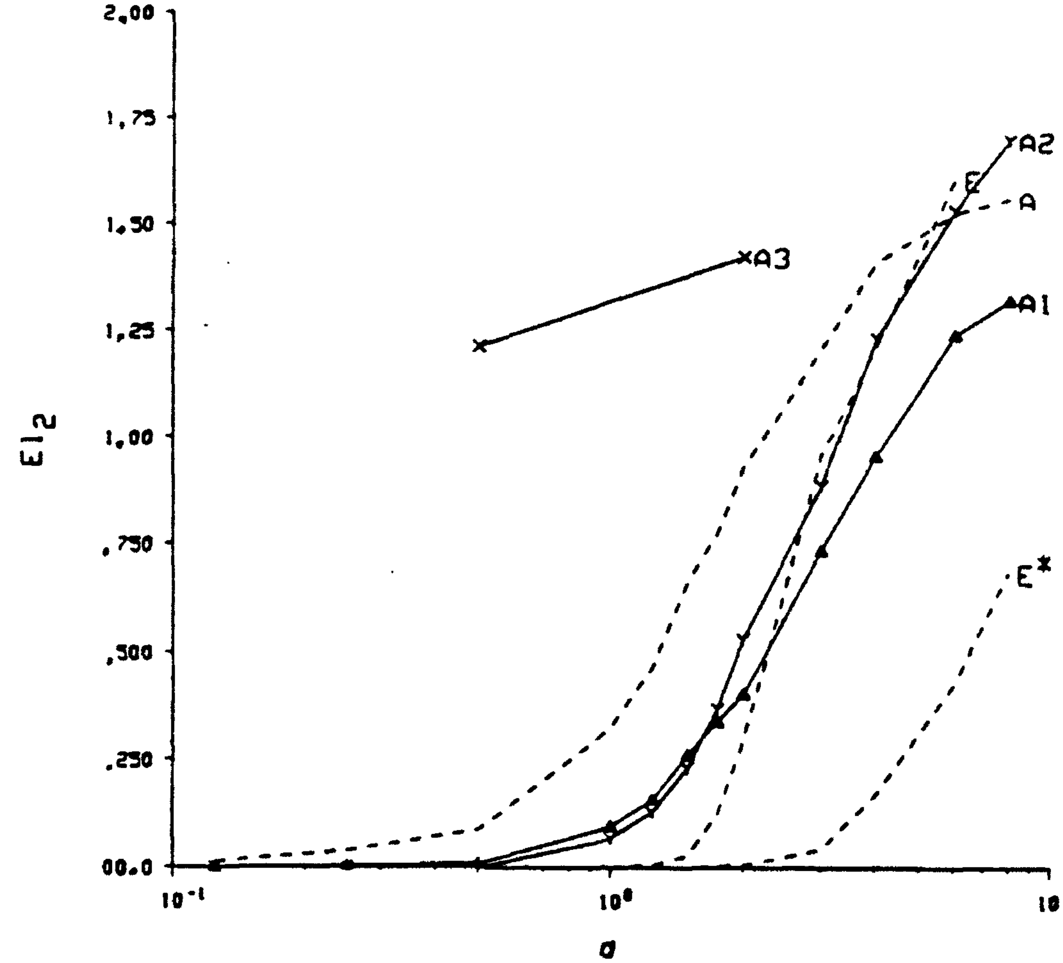

Figure 1: Comparison of 6 estimators on 1D $P\bar{1}$ problems as a function of observation noise standard deviation.

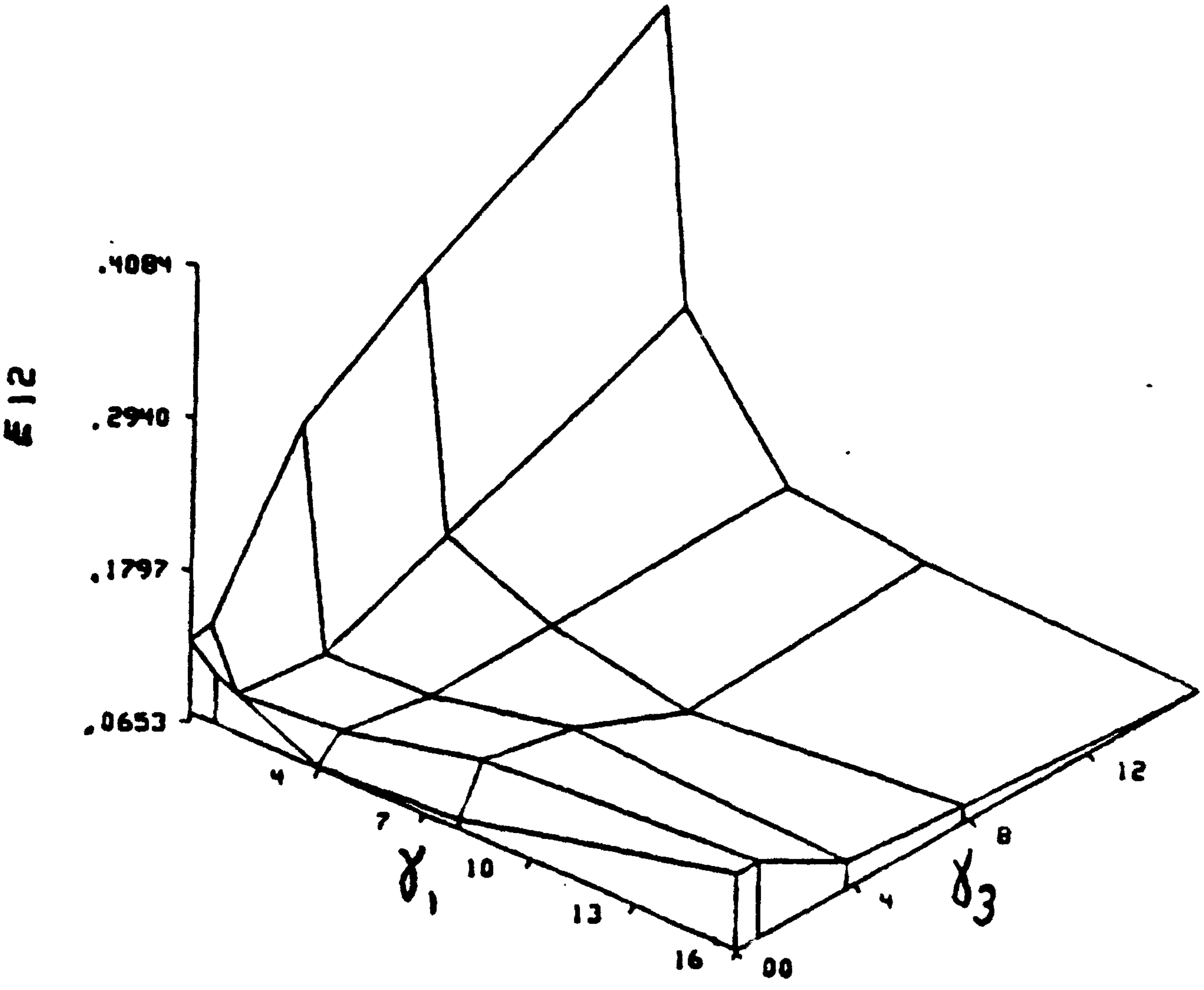

Figure 2: Estimator performance statistics on 3D monoclinic $C2$ problems as a function of the parameters of C.

2.3 Recursive Stochastic Algorithms for Global Optimization in R^d

2.3.1 Introduction

A class of algorithms for finding the global minimum of a smooth function $U(x), x \in \mathbf{R}^d$ are termed *Modified Stochastic Gradient Algorithms.* This section analyzes the convergence of algorithms of the form

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k, \tag{2.63}$$

where $\{\xi_k\}$ is a sequence of $\mathbf{R}^d$-valued random variables, $\{W_k\}$ is a sequence of standard d-dimensional independent Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers with a_k, $b_k \to 0$. An algorithm of this type arises by artificially adding the $b_k W_k$ term (via a Monte Carlo simulation) to the standard stochastic gradient algorithm

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k). \tag{2.64}$$

Algorithms like 2.64 arise in a variety of optimization problems including adaptive filtering, identification and control; here the sequence $\{\xi_k\}$ is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (c.f. [26]). The asymptotic behavior of $\{Z_k\}$ has been much studied. Let S and S^* be the set of local and global minima of $U(\cdot)$, respectively. It can be shown, for example, that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ for k large, and $\{Z_k\}$ is bounded, then $Z_k \to S$ as $k \to \infty$ w.p.1. However, in general $Z_k \not\to S^*$ (unless of course $S = S^*$). The idea behind adding the additional $b_k W_k$ term in 2.63 compared with 2.64 is that if b_k tends to zero slowly enough, then possibly $\{X_k\}$ (unlike $\{Z_k\}$) will avoid getting trapped in a strictly local minimum of $U(\cdot)$ (this is the usual reasoning behind simulated annealing-type algorithms). We shall in fact show that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ and $b_k^2 = B/k \log\log k$ for k large with $B/A > C_0$ (where C_0 is a positive constant which depends only on $U(\cdot)$), and $\{X_k\}$ is tight, then $X_k \to S^*$ as $k \to \infty$ in probability. We also give a condition for the tightness of $\{X_k\}$. We note that the convergence of Z_k to S can be established under very weak conditions on $\{\xi_k\}$ assuming $\{Z_k\}$ is bounded. Here the convergence of X_k to S^* is established under somewhat stronger conditions on $\{\xi_k\}$ assuming that $\{X_k\}$ is tight (which is weaker than boundedness).

2.3.2 Convergence of the Modified Stochastic Gradient Algorithm

The analysis of the convergence of $\{X_k\}$ is usually based on the asymptotic behavior of the associated ordinary differential equation (ODE)

$$\dot{z}t = -\nabla U(z(t)) \tag{2.65}$$

(c.f. [26, 27]). This motivates our analysis of the convergence of $\{X_k\}$ based on the asymptotic behavior of the associated stochastic differential equation (SDE)

$$dY(t) = -\nabla U(Y(t))dt + c(t)dW(t), \tag{2.66}$$

where $W(\cdot)$ is a standard d-dimensional Wiener process and $c(\cdot)$ is a positive function with $c(t) \to 0$ as $t \to \infty$. This is just the diffusion annealing algorithm discussed in [31, Equation (4.3)], with $T(t) = c^2(t)/2$. The asymptotic behavior of $Y(t)$ as $t \to \infty$ has been studied intensively by a number of researchers. In [19, 25], convergence results where obtained by considering a version of 2.66 with a reflecting boundary; in [3], the reflecting boundary was removed. Our analysis of $\{X_k\}$ is based on the analysis of $Y(t)$ developed in [3], where the following result is proved: if $U(\cdot)$ is well-behaved and $c^2(t) = C/\log t$ for t large with $C > C_0$ (the same constant C_0 as above) then $Y(t) \to S^*$ as $t \to \infty$. To see intuitively how $\{X_k\}$ and $Y(\cdot)$ are related, let $t_k = \sum_{n=0}^{k-1} a_n$, $a_k = A/k, b_k^2 = B/k \log\log k$, $c^2(t) = C/\log t$, and $B/A = C$. Note that $b_k \sim c(t_k)\sqrt{a_k}$. Then we should have that

$$\begin{aligned} Y(t_{k+1}) &\simeq Y(t_k) - (t_{k+1} - t_k)\nabla U(Y(t_k)) + c(t_k)(W(t_{k+1}) - W(t_k)) \\ &= Y(t_k) - a_k \nabla U(Y(t_k)) + c(t_k)\sqrt{a_k V_k} \\ &\simeq Y(t_k) - a_k \nabla U(Y(t_k)) + b_k V_k \end{aligned}$$

where $\{V_k\}$ is a sequence of standard d-dimensional independent Gaussian random variables. Hence (for $\{\xi_k\}$ small enough) $\{X_k\}$ and $\{Y(t_k)\}$ should have approximately the same distributions. Of course, this is a heuristic; there are significant technical difficulties in using $Y(\cdot)$ to analyze $\{X_k\}$ because we must deal with long time intervals and slowly decreasing (unbounded) Gaussian random variables.

An algorithm like 2.63 was first proposed and analyzed in [25]. However, the analysis required that the trajectories of $\{X_k\}$ lie within a fixed ball (which as achieved by modifying 2.63 near the boundary of the ball). Hence such a version of 2.63 is only suitable for optimizing $U(\cdot)$ over a compact set. Furthermore the analysis also required ξ_k to be zero in order to obtain convergence. In our first analysis of 2.63 in [16], we also required that the trajectories of $\{X_k\}$ lie in a compact set. However, our analysis did not require ξ_k to be zero, which has important implications when $\nabla U(\cdot)$ is not measured exactly. In our later analysis of 2.63 in [17], we removed the requirement that the trajectories of $\{X_k\}$ lie in a compact set. From our point of view this is the most significant difference between our work in [17] and what is done in [25, 16] (and more generally in other work on global optimization such as [4]): we deal with unbounded processes and establish the convergence of an algorithm which finds a global minimum of a function when it is not specified a priori what bounded region contains such a point.

We now state the simplest result from [17] concerning the convergence of the modified stochastic gradient algorithm 2.63. We will require

$$a_k - \frac{A}{k}, \quad b_k = \frac{\sqrt{b}}{\sqrt{k \log\log k}}, \quad k \text{ large.} \tag{2.67}$$

and the following conditions:

(A1) $U(\cdot)$ is a C^2 function from $\mathbf{R}^d$ to $[0,\infty)$ such that the $S^* = \{x : U(x) \leq U(y)\ \forall\ y\} \neq \emptyset$. (We also require some mild regularity conditions on $U(\cdot)$; see 2.63).

(A2) $\underline{\lim}_{x\to\infty} \frac{|\nabla U(x)|}{|x|} > 0, \overline{\lim}_{x\to\infty} \frac{|\nabla U(x)}{|x|} < \infty$.

(A3) $\lim_{x\to\infty} \left\langle \frac{\nabla U(x)}{|\nabla U(x)|}, \frac{x}{|x|} \right\rangle = 1$

(A4) For $k = 0, 1, \ldots$, let $\mathcal{F}_k$ be the σ-field generated by $X_0, W_0, \ldots, W_{k-1}$, $\xi_0, \ldots, \xi_{k-1}$. There exists an $L \geq 0$, $\alpha > -1$, and $\beta > 0$ such that

$$\mathbf{E}[|\xi_k|^2|\mathcal{F}_k] \leq L a_k^a(|X_k|^2 + 1), \quad |\mathbf{E}[\xi_k|\mathcal{F}_k]| \leq L a_k^\beta(|X_k| + 1) \text{ w.p. } 1$$

and W_k is independent of $\mathcal{F}_k$.

Theorem 2.3.1 *Assume (A1)-(A4) hold. Let $\{X_k\}$ be given by 2.63. Then there exists a constant C_0 such that for $B/A > C_0$*

$$X_k \rightarrow S^* \text{ as } k \rightarrow \infty$$

in probability.

Proof. See [17]. ∎

Remarks:

1. The constant C_0 plays a critical role in the convergence of X_k as $k \rightarrow \infty$ and also $Y(t)$ as $t \rightarrow \infty$. In [3], it is shown that the constant C_0 (denoted there by c_0) has an interpretation in terms of the action functional for a family of perturbed dynamical systems; see [3] for a further discussion of C_0 including some examples.
2. It is possible to modify 2.63 in such a way that only the lower bound and not the upper bound on $|\nabla U(\cdot)|$ in (A2) is needed (see [17]).
3. In [17] we actually separate the problem of convergence of $\{X_k\}$ into two parts: one to establish tightness and another to establish convergence given tightness. This is analogous to separating the problem of convergence of $\{X_k\}$ into two parts: one to establish boundedness and another to establish convergence given boundedness (c.f. [26]). Now in [17] the conditions given for tightness are much stronger than the conditions given for convergence assuming tightness. For a particular algorithm it is often possible to prove tightness directly, resulting in somewhat weaker conditions than those given in [31, Theorem 3.1].

2.3.3 Continuous-State Markov Chain Algorithm

In this section we examine the convergence of a class of continuous-state Markov chain annealing algorithms. Our approach is to write such an algorithm in the form of a modified stochastic gradient algorithm of (essentially) the type considered in Section 2.3.1. A convergence result is obtained for global optimization over all of $\mathbf{R}^d$. Some care is necessary to formulate a Markov chain with appropriate scaling. It turns out that writing the Markov chain annealing algorithm is (essentially) the form 2.63 is rather more complicated than writing standard variations of gradient algorithms which use some type of (possibly noisy) finite difference estimate of $\nabla U(\cdot)$ in the form 2.64 (c.f. [26]). Indeed, to the extend that the Markov chain annealing algorithm uses an estimate of $\nabla U(\cdot)$, it does so in a much more subtle manner than a finite difference approximation.

Although some numerical work has been performed with continuous-state Markov chain annealing algorithm in [23, 36], there has been very little theoretical analysis, and furthermore the analysis of the continuous state case does not follow from the finite state case in a straightforward way (especially for an unbounded state space). The only analysis we are aware of its in [23] where a certain asymptotic stability property is established. Since our convergence results for the continuous state Markov chain annealing algorithm are ultimately based on the asymptotic behavior of the diffusion annealing algorithm, our work demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing.

We shall perform our analysis of continuous state Markov chain annealing algorithms for a Metropolis type chain. We remark that convergence results for other continuous-state Markov chain sampling method-based annealing algorithms (such as the Heat Bath method) can be obtained by a similar procedure. Recall that the 1-step transition probability density for a continuous state Metropolis-type (fixed temperature) Markov chain is given by

$$p(x,y) = q(x,y)s(x,y) + m(x)\delta(y-x)$$

where

$$m(x) = 1 - \int q(x,y)s(x,y)dy$$

and

$$s(x,y) = \exp(-[U(y) - U(x)]_+/T).$$

Here we have dropped the subscript on the weighting factor $s(x,y)$. If we replace the fixed temperature T by a temperature sequence $\{T_k\}$ we get a Metropolis-type annealing algorithm.

Our goal is to express the Metropolis-type annealing algorithm as a modified stochastic gradient algorithm like 2.63 so as to establish its convergence. This leads us to choosing a nonstationary Gaussian transition density

$$q_k(x,y) = \frac{1}{(2\pi b_k^2\sigma^2(x))^{d/2}}\exp(\frac{|y-x|^2}{2b_k^2\sigma^2(x)}) \tag{2.68}$$

$$T_k(x) = \frac{b_k^2\sigma_k^2(x)}{2a_k} \tag{2.69}$$

where $\sigma_k(x) = (\delta_k|x|)v^1, \delta_k \downarrow 0$.

With these choices the Metropolis-type annealing algorithm can be expressed as

$$X_{k+1} = X_k - \alpha_k(\nabla U(X_k) + \xi_k) + b_k\sigma(X_k)W_k \tag{2.70}$$

for appropriately behaved $\{\xi_k\}$. Note that 2.70 is not identical to 2.63 (because $\sigma(x) \not\equiv 1$), but is turns out that Theorem 2.3.1 holds for $\{X_k\}$ generated by either 2.63 or 2.70. We remark that the state dependent term $\sigma(x)$ term in 2.68- 2.69 produces a drift toward the origin proportional to $|x|$, which is needed to establish tightness of the annealing chain.

This discussion leads us to the following continuous-state Metropolis-type annealing algorithm. Let $\mathcal{N}(m,\Lambda)$ denote d-dimensional normal measure with mean m and covariance matrix Λ.

Let $\{X_k\}$ be a Markov chain with 1 step transition probability at time k given by

$$\mathbf{p}(X_{k+1} \in A | X_k = x) = \int_A s_k(x,y)d\mathcal{N}(x, b_k^2\sigma_k^2(x)I)(y) + m_k(x)1_A(x) \tag{2.71}$$

where

$$m_k(x) = 1 - \int s_k(x,y)d\mathcal{N}(x, b_k^2\sigma^2(x)I)(y) \tag{2.72}$$

$$\sigma_k(x) = (a_k^r r|x|)V1 \tag{2.73}$$

$$s_k(x,y) = \exp(-\frac{2a_k[U(y) - U(x)]+}{b_k^2 \quad \sigma_k^2(x)}) \tag{2.74}$$

A convergence result similar to the previous theorem can be proved for the Metropolis type annealing algorithms (c.f. [18]).

2.4 References

[1] G. Bricogne, "Maximum Entropy and the Foundations of Direct Methods," *Acta Crystallographica A*, 40, 1984, 410–445.

[2] G. Bricogne, "A Bayesian Statistical Theory of the Phase Problem. I. Multichannel Maximum Entropy Formalism for Constructing Generalized Joint Probability Distributions of Structure Factors," *Acta Crystallographica A*, 44, 1988, 517–545.

[3] T.S. Chiang, C.R. Hwang, and S.J. Sheu, "Diffusion for Global Optimization in $\mathbf{R}^n$," *SIAM J. Control and Optimization*, 25, 1987, 737-752.

[4] L.C.W. Dixon and G.P. Szego, *Towards Global Optimisation*, North Holland, Amsterdam, 1978.

[5] P.C. Doerschuk, "Direct Methods in Single Crystal X-Ray Crystallography—Part 2: Generalizations of Mean Field Theory," Rep. LIDS-WP-1916, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1989.

[6] P.C. Doerschuk, "Adaptive Bayesian Signal Reconstruction with a priori Model Implementation and Synthetic Examples for X-Ray Crystallography," *J. Optical Society of America A*, 8, 1991, 1222–1232.

[7] P.C. Doerschuk, "Bayesian Signal Reconstruction from Fourier Transform Magnitude and X-Ray Crystallography," in S.-S. Chen, ed., *Stochastic and Neural Methods in Signal Processing, Image Processing, and Computer Vision*, San Diego, California, 24–26 July 1991, SPIE The International Society for Optical Engineering, vol. 1569, 70–79.

[8] P.C. Doerschuk, "Bayesian Signal Reconstruction, Markov Random Fields, and X-Ray Crystallography," *J. Optical Society of America A*, 8, 1991, 1207–1221.

[9] P.C. Doerschuk, "Multidimensional Bayesian Signal Reconstruction from Fourier Transform Magnitude and X-Ray Crystallography," in *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*, Lake Placid, New York, 23–25 September 1991, IEEE Signal Processing Society, 6–8.

[10] P.C. Doerschuk, "Bayesian Phase Retrieval for X-Ray Crystallography," in M.A. Fiddy, ed., *Inverse Problems in Scattering and Imaging*, San Diego, California, 20–22 July 1992, SPIE The International Society for Optical Engineering, vol. 1767.

[11] P.C. Doerschuk, "Bayesian Reconstruction of Signals Invariant under a Space Group Symmetry from Fourier Transform Magnitudes and X-Ray Crystallography," submitted to *IEEE Trans. Image Processing.*

[12] P.C. Doerschuk. "Bayesian Signal Reconstruction from Fourier Transform Magnitude in the Presence of Symmetries and X-Ray Crystallography." Rep. TR-EE-92-16, School of Electrical Engineering, Purdue University, 1992.

[13] P.C. Doerschuk, "Signal Reconstruction from Fourier Transform Magnitude using Markov Random Fields in X-Ray Crystallography," in *Proceedings: 1992 International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, California. 23–26 March 1992. IEEE Signal Processing Society, vol. 4, 141–144.

[14] P.C. Doerschuk, "Use of Hard Constraints for Bayesian Reconstruction of Symmetric Signals from Fourier Transform Magnitudes and X-Ray Crystallography," submitted to *IEEE Trans. Image Processing.*

[15] P.C. Doerschuk, "X-Ray Crystallography as a Baysian Signal Reconstruction Problem," in *Signal Recovery and Synthesis IV*, New Orleans, Louisiana, 14–16 April 1992, Optical Society of America. Technical Digest Series vol. 11, 28–30.

[16] S.B. Gelfand and S.K. Mitter, "Simulated Annealing-Type Algorithms for Multivariate Optimization," *Algorithmica*, 6, 1991, 419–436.

[17] S.B. Gelfand and S.K. Mitter, "Recursive Stochastic Algorithms for Global Optimization in $\mathbf{R}^d$," *SIAM J. Control and Optimization*, 29, 1991, 999–1018.

[18] S.B. Gelfand and S.K. Mitter, "Metropolis-Type Annealing Algorithms for Global Optimization in $\mathbf{R}^d$," *SIAM J. Control and Optimization*, 31, 1993, 111-131.

[19] S. Geman and C.R. Hwang, "Diffusions for Global Optimization," *SIAM J. Control and Optimization*, 24, 1986, 1031–1043.

[20] C. Giacovazzo, *Direct Methods in Crystallography*, Academic Press, London, 1980.

[21] T. Hahn, ed., *International Tables for X-Ray Crystallography, Volume A: Space Group Symmetry*, D. Reidel Publishing Company, Dordrecht-Boston-Lancaster-Tokyo, 1987.

[22] IMSL, Inc., *Users Manual, IMSL MATH/LIBRARY*, Version 1.1, IMSL, Inc., 1989.

[23] F.J. Jeng and J.W. Woods, "Simulated Annealing in Compound Gaussian Random Fields," *IEEE Trans. Information Theory*, IT-36, 1990, 94–107.

[24] W. Kim and M.H. Hayes, "The Phase Retrieval Problem in X-Ray Crystallography," in *Proceedings of ICASSP*, 1991, 1765–1768.

[25] H J. Kushner, "Asymptotic Global Behavior for Stochastic Approximation and Diffusions with Slowly Decreasing Noise Effects: Global Minimization Via Monte Carlo," *SIAM J. Applied Mathematics*, 47, 1987, 169–185.

[26] H.J. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Germany, 1987.

[27] L. Ljung, "Analysis of Recursive Stochastic Algorithms," *IEEE Trans. Automatic Control*, AC-22, 1977, 551–575.

[28] C.J. Masreliez, "Approximate Non-Gaussian Filtering with Linear State and Observation Relations," *IEEE Trans. Automatic Control*, AC-20, 1975, 107-110.

[29] C.J. Masreliez and R.D. Martin, "Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter," *IEEE Trans. Automatic Control*, AC-22, 1977, 361-371.

[30] R.P. Millane, "Phase Retrieval in Crystallography and Optics," *J. Optical Society of America A*, 7, 1990, 394–411.

[31] S.K. Mitter, "Modelling and Estimation for Random Fields," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2167, 1992.

[32] S.K. Mitter and I.C. Schick, "Point Estimation, Stochastic Approximation, and Robust Kalman Filtering," in A. Isidori and T.J. Tarn, eds., *Systems, Models and Feedback: Theory and Applications*, Birkhauser, Boston-Basel-Berlin, 1992, 127–151.

[33] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1988.

[34] I.C. Schick, *Robust Recursive Estimation of the State of a Discrete-Time Stochastic Linear Dynamic System in the Presence of Heavy-Tailed Observation Noise*. Rep. LIDS-TH-1975, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1990.

[35] I.C. Schick and S.K. Mitter, "Robust Recursive Estimation in the Presence of Heavy-Tailed Observation Noise," *Ann. Stat.*, to appear.

[36] T. Simchony, R. Chellappa and Z. Lichtenstein, "Relaxation Algorithms for MAP Estimation of Grey-Level Images with Multiplicative Noise," *IEEE Trans. Information Theory*, IT-36, 1990, 608–613.

3 Staff Supported

3.1 Graduate Students Supported

Mohamed Akra

Stefano Casadei

Irvin C. Schick

3.2 Post-Doctoral Fellows Supported

Peter C. Doerschuk

Denis Mustafa

3.3 Faculty Supported

Sanjoy K. Mitter

4 Invited Presentations

"Variational Problems with Free Discontinuities and Nonlinear Diffusions: Applications in Image Analysis," Systems Research Center, University of Maryland, April 1990.

"Variational Problems with Free Discontinuities and Nonlinear Diffusions: Applications in Image Analysis," Boston University, April 1990.

"Structure and Analysis of Complex Signals," plenary address, 25th Anniversary Symposium. Centre International des Sciences Mechaniques. Udine, Italy, May 1990.

"Variational Problems with Free Discontinuities and Nonlinear Diffusions," 8th Army Conference on Applied Mathematics and Computing, Cornell University, June 1990.

"Early Vision Problems and Nonlinear Diffusions," invited keynote address, Workshop on Signal Processing, Communications and Networking, Indian Institute of Science, Bangalore, India, July 1990.

"Recursive Algorithms," University of Minnesota, November 1990.

"Variational Methods for Image Segmentation," Institute for Mathematics and its Applications, University of Minnesota, April 1991.

"Stochastic Recursive Algorithms for Global Optimization in $\mathbf{R}^d$," Rutgers Workshop on Applied Stochastic Analysis, May 1991.

5 Publications

Dembo, A. and O. Zeitouni, "Large Deviations and Applications: the Finite-Dimensional Case," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2030, April 1991.

Doerschuk, Peter C., "Adaptive Bayesian Signal Reconstruction with a priori Model Implementation and Synthetic Examples for X-Ray Crystallography," Laboratory for Information and Decision Systems, Technical Report LIDS-P-1993; *Journal of the Optical Society of America A*, vol.8, 1991, 1222–1232.

Doerschuk, Peter C., "Bayesian Signal Reconstruction, Markov Random Fields, and X-Ray Crystallography," Laboratory for Information and Decision Systems, Technical Report LIDS-P-1992; *Journal of the Optical Society of America A*, vol.8, 1991, 1207–1221.

Doerschuk, Peter C., "Direct Methods in Single Crystal X-Ray Crystallography, Part I: Introduction and Markov Random Field Models," Laboratory for Information and Decision Systems, Technical Report LIDS-WP-1915, May 1990.

Doerschuk, Peter C., "Direct Methods in Single Crystal X-Ray Crystallography, Part II: Generalization of Mean Field Theory," Laboratory for Information and Decision Systems, Technical Report LIDS-WP-1916, October 1989.

Doerschuk, Peter C., "Direct Methods in Single Crystal X-Ray Crystallography, Part III: the Spherical Approximation and Small- Noise Asymptotics," Laboratory for Information and Decision Systems, Technical Report LIDS-WP-1917, June 1990.

Gelfand, S.B. and S.K. Mitter, "Recursive Stochastic Algorithms for Global Optimization in $\mathbf{R}^d$," Laboratory for Information and Decision Systems, Technical Report LIDS-P-1937; *SIAM Journal on Control and Optimization*, vol.29, 1991, 999–1018.

Gelfand, S.B. and S.K. Mitter, "Metropolis-Type Annealing Algorithms for Global Optimization in $\mathbf{R}^d$," Laboratory for Information and Decision Systems, Technical Report LIDS-P-1977; *SIAM Journal on Control and Optimization*, vol.31, 1993, 111-131.

Gelfand, S.B. and S.K. Mitter, "On Sampling Methods and Annealing Algorithms," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2008; in *Markov Random Fields: Theory and Applications*, R. Chellappa and A. Jain, eds., New York: Academic Press, 1991.

Gelfand, S.B. and S.K. Mitter, "Weak Convergence of Markov Chain Sampling Methods and Annealing Algorithms to Diffusions," *Journal of Optimization Theory and Applications*, vol.63 (1991).

Gelfand, S.B. and S.K. Mitter, "Simulated Annealing-Type Algorithms for Multivariate Optimization," *Algorithmica*, 6, 1991, 419–436.

Kulkarni, S.R. and S.K. Mitter, "Some Discrete Approximations to a Variational Method for Image Segmentation," Laboratory for Information and Decision Systems Technical Report LIDS-P-2014, January 1991.

Mitter, S.K., "Markov Random Fields, Stochastic Quantization and Image Analysis," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2013; in *The State of the Art in Applied and Industrial Mathematics*, R. Spigler, ed., Dordrecht, The Netherlands: Kluwer, 1990.

Mitter, S.K., "Modelling and Estimation for Random Fields," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2167, 1992.

Mitter, S.K. and I.C. Schick, "Point Estimation, Stochastic Approximation, and Robust Kalman Filtering," in A. Isidori and T.J. Tarn, eds., *Systems, Models and Feedback: Theory and Applications*, Birkhauser, Boston-Basel-Berlin, 1992, 127–151.

Mustafa, D. and M.A. Dahleh, "Reduced-Order Control of Systems for which Balanced Truncation is Hankel-Norm Optimal," Laboratory for Information and Decision Systems Technical Report LIDS-P-2034, May 1991; submitted to *IEEE Transactions on Automatic Control.*

Mustafa, D. and K. Glover, *Minimum Entropy H_∞ Control*, Laboratory for Information and Decision Systems, Technical Report LIDS-P-1946, 1989; *Lecture Notes in Control and Information Science*, Volume 146, Springer-Verlag, 1990.

Mustafa, D. and D.S. Bernstein, "LQG Cost Bounds in Discrete-Time H_2/H_∞ Control," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2023, 1990; *Transactions of the Institute of Measurement and Control*, to appear.

Mustafa, D., "Combined H_∞/LQG Control via the Optimal Projection Equations: on Minimizing the LQG Cost Bound," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2024, 1991; *International Journal of Robust and Nonlinear Control*, to appear.

Mustafa, D. and K. Glover, "Controller Reduction by H_∞ Balanced Truncation," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2025; *Int. J. Robust and Nonlinear Control*, to appear.

Mustafa, D., "A Class of Systems for which Balanced Truncation is Hankel-Norm Optimal," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2030, 1991; *30th IEEE Conference on Decision and Control*, 1991.

Richardson, T.J. and S.K. Mitter, "Scaling results for the variational approach to edge detection," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2016, 1991; submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Schick, Irvin C., "Robust Recursive Estimation of the State of a Discrete-Time Stochastic Linear Dynamic System in the Presence of Heavy-Tailed Observation Noise," Ph.D. thesis, Department of Mathematics, Massachusetts Institute of Technology, May 1989; rep. Laboratory for Information and Decision Systems Technical Report LIDS-TH-1975, 1990.

Schick, I.C. and S.K. Mitter, "Robust Recursive Estimation in the Presence of Heavy-Tailed Observation Noise," Laboratory for Information and Decision Systems, Technical Report LIDS-P-2166; *Ann. Stat.*, to appear.